RESPFOCUS: LEARNING TEMPORAL EVIDENCE FOR WEAKLY SUPERVISED RESPIRATORY SOUND CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Respiratory auscultation provides a low-cost means of assessing pulmonary function, yet transient crackles and sustained wheezes can be difficult to recognize in noisy clinical recordings. Conventional deep learning classifiers often operate as black boxes, offering limited insight into the temporal acoustic evidence supporting their predictions. We present RespFocus, a weakly supervised framework that uses multiple instance learning (MIL) to learn inspectable temporal evidence from cycle-level labels and integrate it into classification. We also introduce PedResp2026, a multicenter pediatric respiratory sound cohort collected in routine clinical environments with adjudicated cycle-level annotations. In classification experiments, RespFocus achieves four-class Scores of 69.55% on public ICBHI 2017 and 72.91% on PedResp2026. Evidence diagnostics and matched temporal interventions support the functional contribution of the learned evidence to model predictions. Together, these findings highlight RespFocus's potential as a tool for automated respiratory sound diagnosis.

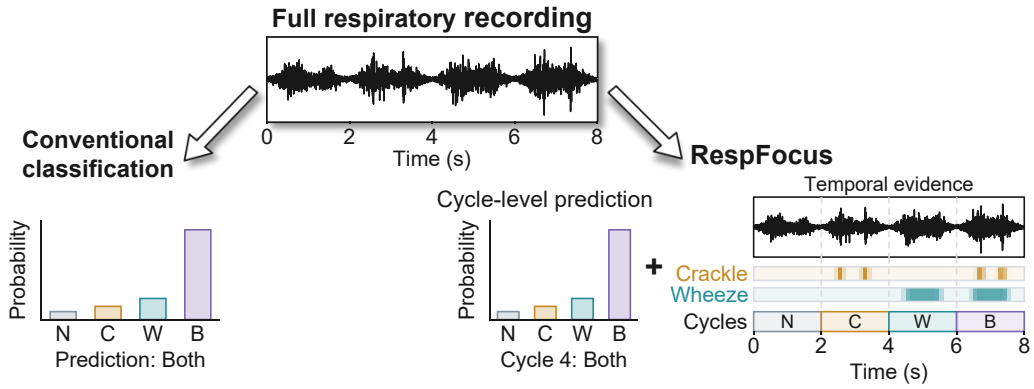


Figure 1: Illustrative comparison of conventional cycle classification (left) and RespFocus (right) with cycle intervals and temporal evidence. Amber/teal denote crackle/wheeze evidence. N/C/W/B denote Normal/Crackles/Wheezes/Both.

1 INTRODUCTION

Respiratory auscultation provides a low-cost means of assessing respiratory sounds in routine care (Rocha et al., 2019; Bardou et al., 2018; Gairola et al., 2021). Crackles typically appear as brief, discontinuous transients, whereas wheezes exhibit a more sustained tonal structure (Pasterkamp et al., 2016). Automated respiratory sound classification remains challenging because these sounds differ in duration and can be obscured by speech, motion, and sensor-contact noise.

A respiratory cycle may contain crackles, wheezes, or both, and the two subtypes need not occur at the same time. Recognizing co-occurrence therefore requires gathering evidence for each subtype across the cycle. Keeping this evidence linked to the recording would also allow a listener to revisit the corresponding portions of the sound when reviewing a prediction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Pretrained audio representations have improved respiratory sound classification (Bae et al., 2023; He et al., 2024; Zhang et al., 2024), providing a starting point for learning such evidence. However, the annotations used here specify cycle boundaries and sound categories without marking individual events. Under this weak supervision, the model must learn which temporal responses are useful by aggregating them into a prediction and comparing it with the cycle label.

Averaging across a cycle can dilute a brief crackle among less informative positions, while a single maximum captures only one response from distributed wheeze activity. A common aggregation rule therefore imposes the same temporal preference on acoustically different subtypes. This motivates separate rules that preserve crackle peaks and combine multiple wheeze responses.

We present RespFocus, a weakly supervised framework that uses multiple instance learning (MIL) (Dietterich et al., 1997; Ilse et al., 2018) to learn inspectable temporal evidence from cycle-level labels and integrate it into classification. Each annotated cycle forms a bag of temporal instances. Building on interpretable time-series MIL (Early et al., 2024), it combines OPERA-CT and Audio Spectrogram Transformer (AST) representations (Zhang et al., 2024; Gong et al., 2021). A shared abnormal-presence head guides local refinement in separate subtype branches. Crackle max pooling preserves peak responses, while wheeze top-k pooling combines multiple high-scoring positions. The two subtype decisions compose the four cycle categories and represent co-occurrence within a cycle. Each prediction is accompanied by shared and subtype-specific evidence indexed to the supplied interval (Figure 1).

To evaluate this approach alongside ICBHI 2017, we introduce PedResp2026, a new multicenter pediatric respiratory sound dataset collected during routine care. It contains 493 recordings from 137 children at seven hospitals, with expert-adjudicated cycle labels. Patient-disjoint partitions support evaluation on held-out children from the collection. The new cohort complements the public benchmark with recordings from pediatric care and provides a resource for learning temporal evidence under weak supervision.

RespFocus improves classification on both datasets. Progressive ablations examine how performance changes as the architecture is assembled, with a controlled comparison isolating the pooling design. Matched feature-space suppression and retention test whether high-ranked evidence contributes to predictions. These experiments evaluate the classification outcome, the design choices behind it, and the role of the selected evidence. Our contributions are:

1. A weakly supervised framework that uses MIL to learn inspectable temporal evidence from cycle-level labels and integrate it into respiratory sound classification.
2. A morphology-aware aggregation and decoding design that uses crackle max pooling and wheeze top-k pooling to reflect their temporal differences, then composes the two subtype decisions into four cycle categories.
3. PedResp2026, a new multicenter pediatric respiratory sound dataset with 493 recordings from 137 children at seven hospitals, expert-adjudicated cycle labels, and patient-disjoint evaluation partitions.
4. Evaluation on ICBHI 2017 and PedResp2026, with four-class Scores of 69.55% and 72.91%, respectively. Architectural ablations and matched feature-space interventions on ICBHI 2017 connect the classification gains to the proposed design and the evidence used for prediction.

2 Related Work

Weakly supervised temporal learning. MIL learns from bag labels without instance annotations (Dietterich et al., 1997). Respiratory cycle annotations follow this structure, with an acoustic label for each cycle and latent locations for the supporting sounds. Attention-based MIL learns instance aggregation (Ilse et al., 2018). MILLET extends this approach to time-series classification and local explanations (Early et al., 2024). Both make aggregation an explicit link between local responses and a sequence-level decision. Weakly supervised action localization separates foreground activity from background (Paul et al., 2018; Nguyen et al., 2019; Ren et al., 2023), while sound event detection addresses event duration and acoustic background (Dinkel et al., 2021; Xin et al., 2023). RespFocus uses cycle labels to learn evidence within supplied intervals. Its subtype-specific pooling encodes

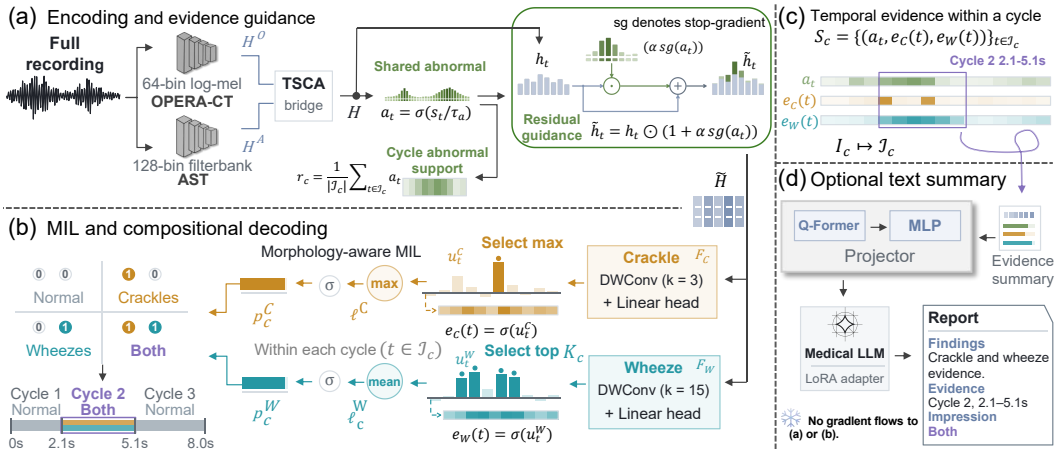


Figure 2: Illustrative RespFocus overview. (a) Dual-stream fusion and guidance from shared abnormal evidence. (b) Subtype MIL pooling and compositional decoding. (c) Shared abnormal, crackle, and wheeze evidence. (d) Optional text summaries with temporal references.

the different temporal preferences of crackles and wheezes, and its compositional decoder preserves their co-occurrence within a bag.

Respiratory sound representations and explanations. Convolutional and transformer models support respiratory sound classification (Bardou et al., 2018; Gairola et al., 2021; Bae et al., 2023; He et al., 2024). Supervised audio pretraining, masked modeling, and acoustic tokenization provide transferable features (Kong et al., 2020; Gong et al., 2022; Huang et al., 2022; Chen et al., 2023). OPERA specializes pretraining to respiratory audio (Zhang et al., 2024). RespFocus builds on OPERA-CT and AST representations by guiding, refining, and pooling their temporal features for the two sound subtypes.

Prototype and example-based approaches make predictions interpretable through reference examples (Ren et al., 2022; Chang et al., 2022). RespFocus exposes temporal evidence within the input cycle and tests its contribution through interventions on the model's representations. The shared evidence describes abnormal support, and the subtype scores expose the responses pooled into the cycle-level logits. This makes the evidence accessible along the same path that produces the prediction.

3 Methodology

Given a recording x and annotated cycle intervals $\{I_c\}_{c=1}^{N}$, RespFocus returns $\{(I_c, \hat{y}_c, S_c)\}_{c=1}^{N}$: the supplied intervals, predicted categories, and indexed temporal evidence. Each interval defines an instance bag, with predicted category $\hat{y}_c \in \{\text{Normal}, \text{Crackles}, \text{Wheezes}, \text{Both}\}$. Both denotes crackles and wheezes within the same cycle, whether or not their events overlap.

The model fuses OPERA-aligned temporal features with AST context, uses shared abnormal evidence to guide subtype refinement, and pools the subtype logits within each cycle (Figure 2). The shared head modulates temporal features according to abnormal support. The subtype branches refine these features into crackle and wheeze responses, which cycle pooling converts into the final decision. The optional text interface verbalizes the acoustic outputs. Its implementation and illustrative cases appear in Appendices B.3 and I.

3.1 Dual-stream Temporal Encoding

OPERA-CT encodes a 64-bin log-mel spectrogram, while AST encodes a normalized 128-bin filterbank (Zhang et al., 2024; Gong et al., 2021). Writing the frontends as $\mathcal{M}_{64}$ and $\mathcal{F}_{128}$, we obtain

$$\begin{aligned} X^O &= \mathcal{M}_{64}(x), \quad H^O = f_O(X^O) \in \mathbb{R}^{T\times 768}, \\ X^A &= \mathcal{F}_{128}(x), \quad H^A = f_A(X^A) \in \mathbb{R}^{T_A\times 768}. \end{aligned} \tag{1}$$

Here, T is the length of the OPERA-aligned temporal sequence, and T_A is the number of AST tokens.

Temporal-spectral cross-attention (TSCA) projects both outputs to $d_b = 1024$ and queries AST features with the OPERA-aligned sequence (Vaswani et al., 2017),

$$\begin{aligned} Z^O &= H^O W_O, \qquad Z^A = H^A W_A, \\ \hat{Z}^O &= \mathrm{LN}\big(Z^O + \mathrm{MHA}(Q = Z^O, K = Z^A, V = Z^A)\big). \end{aligned} \tag{2}$$

A residual gate and output projection produce the shared temporal features,

$$\begin{aligned} g^O &= \sigma\Big(\phi_O(\hat{Z}^O)\Big), \\ \bar{Z}^O &= \hat{Z}^O \odot (0.1 + 0.9 g^O), \\ H &= \mathrm{MLP}_{\mathrm{out}}(\bar{Z}^O) = [h_1, \ldots, h_T] \in \mathbb{R}^{T \times D}. \end{aligned} \tag{3}$$

Here, $D = 512$. Each supplied interval I_c maps to a nonempty token set $\mathcal{I}_c \subseteq \{1, \ldots, T\}$ using the nearest OPERA-token center times. Features are computed before cycle grouping. AST retains at most the first 512 input frames, so its context covers the initial segment of longer recordings. OPERA supplies the temporal indexing. Appendix C details the frontends.

3.2 Shared Abnormal-presence Evidence and Subtype Refinement

For each fused token h_t, a shared head combines local evidence and an auxiliary abnormal cue into a logit s_t (Appendix B.1). Its score modulates the feature through residual guidance,

$$a_t = \sigma(s_t/\tau_a), \qquad \tilde{h}_t = h_t \odot \big(1 + \alpha\ \mathrm{sg}(a_t)\big). \tag{4}$$

Here, $\tau_a > 0$ is the sigmoid temperature, α is the guidance strength, and sg stops gradients through the multiplier, broadcast across channels. The direct feature path remains trainable, and the shared head receives a separate abnormal-presence objective (Section 3.4).

Inspired by convolutional local modeling (Gulati et al., 2020; Peng et al., 2022), we refine the guided sequence $\tilde{H} = [\tilde{h}_1, \ldots, \tilde{h}_T]$ with separate subtype blocks, using a shorter kernel for crackles and a broader kernel for wheezes. Each block is followed by a linear subtype head,

$$\begin{aligned} H^C &= F_C(\tilde{H}; k = 3), \qquad u_t^C = w_C^\top h_t^C + b_C, \\ H^W &= F_W(\tilde{H}; k = 15), \quad u_t^W = w_W^\top h_t^W + b_W. \end{aligned} \tag{5}$$

Kernel sizes are in tokens and express relative temporal preferences. The subtype scores $e_C(t) = \sigma(u_t^C)$ and $e_W(t) = \sigma(u_t^W)$ form the indexed evidence collection

$$S_c = \big\{\big(a_t, e_C(t), e_W(t)\big)\big\}_{t \in \mathcal{I}_c}. \tag{6}$$

Shared evidence a_t guides refinement, while $e_C(t)$ and $e_W(t)$ expose the subtype responses whose logits are pooled for classification. All three scores retain correspondence with the recording.

3.3 Morphology-aware MIL and Compositional Decoding

Crackle pooling selects the largest logit, while wheeze pooling averages multiple high-scoring positions:

$$\begin{aligned} \ell_c^C &= \max_{t \in \mathcal{I}_c} u_t^C, \qquad K_c = \max\big(1, \lceil \rho_W |\mathcal{I}_c| \rceil\big), \quad \rho_W = 0.35, \\ \ell_c^W &= \frac{1}{K_c} \sum_{t \in \mathrm{TopK}_{K_c}(\mathcal{I}_c; u^W)} u_t^W. \end{aligned} \tag{7}$$

TopK_{K_c} selects the K_c largest wheeze logits, whose positions can be noncontiguous. The count K_c differs from the convolutional kernel size k. Building on MIL aggregation (Ilse et al., 2018; Early et al., 2024), we use asymmetric pooling to preserve peak crackle responses while combining

distributed wheeze evidence. Both operations produce cycle-level logits from the same temporal evidence exposed for inspection.

The pooled logits yield two subtype probabilities, which are thresholded independently using validation-selected thresholds,

$$p_c^q = \sigma(\ell_c^q), \qquad \hat{b}_c^q = \mathbf{1}[p_c^q > \theta_q], \qquad q \in \{C, W\}. \tag{8}$$

The ordered pair of crackle and wheeze decisions determines the category,

$$\hat{y}_c = \begin{cases} \text{Normal}, & (\hat{b}_c^C, \hat{b}_c^W) = (0,0), \\ \text{Crackles}, & (\hat{b}_c^C, \hat{b}_c^W) = (1,0), \\ \text{Wheezes}, & (\hat{b}_c^C, \hat{b}_c^W) = (0,1), \\ \text{Both}, & (\hat{b}_c^C, \hat{b}_c^W) = (1,1). \end{cases} \tag{9}$$

The decoder assigns Both when both subtype decisions are positive. Shared abnormal evidence guides feature refinement upstream of this decision. Threshold grids and selected pairs appear in Appendix C.

3.4 Training Objective

Let $y_c^C, y_c^W \in \{0,1\}$ encode the presence of each subtype, and let $y_c^A = \mathbf{1}[y_c^C \vee y_c^W]$ denote abnormal presence. The primary objective applies asymmetric loss (ASL) to the MIL-pooled subtype logits (Ridnik et al., 2021),

$$\mathcal{L}_{\text{pair}} = \frac{1}{N} \sum_{c=1}^{N} \text{ASL}\big((\ell_c^C, \ell_c^W), (y_c^C, y_c^W)\big). \tag{10}$$

The auxiliary objective trains the shared head to recognize abnormal support across both subtypes. It averages the shared evidence logits within each cycle and applies a binary abnormal-presence loss,

$$\begin{aligned} \bar{s}_c &= \frac{1}{|\mathcal{I}_c|} \sum_{t \in \mathcal{I}_c} s_t, \\ \mathcal{L}_{\text{abn}} &= \frac{1}{N} \sum_{c=1}^{N} \text{BCEWithLogits}(\bar{s}_c, y_c^A). \end{aligned} \tag{11}$$

The acoustic model is optimized with

$$\mathcal{L} = \mathcal{L}_{\text{pair}} + \omega_{\text{abn}} \mathcal{L}_{\text{abn}}, \qquad \omega_{\text{abn}} = 1. \tag{12}$$

Both labels activate both subtype targets, while Normal labels set both to zero. Each subtype branch can therefore learn from cycles containing its subtype alone or together with the other subtype. The losses supervise temporal responses through cycle aggregation, allowing evidence learning without event-level annotations.

For evidence analysis and presentation, we separately summarize the shared scores as

$$r_c = \frac{1}{|\mathcal{I}_c|} \sum_{t \in \mathcal{I}_c} a_t. \tag{13}$$

Thus, analysis uses the mean probability r_c, whereas training uses the mean logit $\bar{s}_c$.

4 EXPERIMENTAL SETUP

4.1 DATASETS

We evaluate respiratory sound classification on ICBHI 2017 (Rocha et al., 2019) and PedResp2026. Our fixed ICBHI partition uses 777/90/53 recordings and 5,815/657/426 cycles for training, validation, and testing. Test recordings follow the OPERA held-out list.

PedResp2026 is a newly collected pediatric cohort comprising 493 recordings from 137 children at seven hospitals, with nine device identifiers. Children aged 0.1–13 years were recorded in routine clinical rooms using an eight-site thoracic protocol. Two physicians annotated cycles, with adjudication by an additional physician from an eight-physician panel. The patient-disjoint training, validation, and test sets contain 109/14/14 children, 394/49/50 recordings, and 899/119/117 respiratory cycles, respectively. Appendix A provides class counts, acquisition procedures, ethics information, and the controlled-access release plan.

4.2 BASELINES AND IMPLEMENTATION

We compare with respiratory sound classifiers using convolutional, transformer, contrastive, masked-audio, and adaptation-based approaches (Moummad & Farrugia, 2023; Bae et al., 2023; Kim et al., 2023; 2024b;a; Niizumi et al., 2024; He et al., 2024; Dong et al., 2025a;b). We also consider general MIL baselines for time-series classification (Early et al., 2024; Chen et al., 2024), anomaly detection (Jang & Kwon, 2025), and instance aggregation (Tang et al., 2024), alongside an interpretable mixture-of-experts model (Xie et al., 2026), for adaptation to respiratory sound classification.

OPERA-CT uses 16-kHz audio, 64 mel bins, a 2,048-sample FFT/window, and a 512-sample hop. AST uses a normalized 128-bin Kaldi filterbank, a 10-ms frame shift, and 512 input frames. We train with AdamW and cosine learning-rate decay over five independent seeds. Checkpoints and subtype thresholds are selected on validation data and fixed for testing. The wheeze pooling ratio is $\rho_W = 0.35$, and the crackle and wheeze kernels are 3 and 15 tokens. Appendix C gives checkpoints, optimization settings, and threshold grids.

4.3 EVALUATION METRICS

Classification. We use the ICBHI challenge metric definitions (Rocha et al., 2019) on the study partitions above. Let M_{ij} count cycles with reference class i and predicted class j, where N, C, W, B denote Normal, Crackles, Wheezes, and Both. Specificity (Sp) measures correct normal classification. Sensitivity (Se) measures correct abnormal-subtype classification, penalizing both missed abnormalities and subtype confusions:

$$\begin{aligned} \mathrm{Sp} &= \frac{M_{NN}}{\sum_j M_{Nj}}, \\ \mathrm{Se} &= \frac{M_{CC} + M_{WW} + M_{BB}}{\sum_{i\in\{C,W,B\}} \sum_j M_{ij}}, \\ \mathrm{Score} &= \tfrac{1}{2}(\mathrm{Sp} + \mathrm{Se}). \end{aligned} \tag{14}$$

We report Sp, Se, and Score as percentages, alongside macro-AUROC and macro-F1. Appendix D provides additional metric details and class-wise results.

Temporal evidence. We evaluate two properties of the shared evidence. Abnormal-presence AUROC uses r_c to rank cycles labeled Normal versus Abnormal. CycleHitRate measures the fraction of abnormal cycles containing at least one token above a validation-selected evidence threshold,

$$\mathrm{CycleHitRate} = \frac{1}{|\mathcal{C}_{\mathrm{abn}}|} \sum_{c\in\mathcal{C}_{\mathrm{abn}}} \mathbf{1}\left[\max_{t\in\mathcal{I}_c} a_t \geq \tau_{\mathrm{hit}}\right]. \tag{15}$$

Here, $\mathcal{C}_{\mathrm{abn}}$ contains abnormal cycles. We test functional relevance by suppressing or retaining selected tokens under matched feature-space budgets, including random, low-evidence, and energy-matched controls (Appendix H).

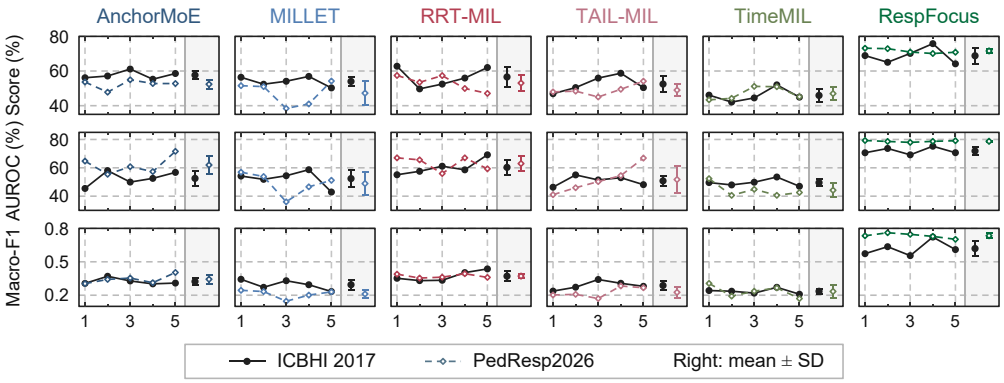


Figure 3: Five-run results by metric (rows) and method (columns) for ICBHI 2017 (solid black lines, filled circles) and PedResp2026 (colored dashed lines, open diamonds). Right summaries show mean ± sample SD. Run indices do not imply paired initialization across methods.

5 Results

5.1 Classification performance

RespFocus achieves the highest mean four-class Score among the compared methods on both datasets (Table 1). Scores of 69.55% on ICBHI and 72.91% on PedResp2026 exceed the strongest baselines, MVST (66.56%) and Patch-Mix CL (68.41%), by 2.99 and 4.50 percentage points. RespFocus also achieves the highest sensitivity, improving by 4.92 and 16.06 points over these baselines. On PedResp2026, sensitivity rises from 52.12% to 68.18%, showing stronger abnormal-subtype recognition. Appendix E reports standard deviations and binary results.

Table 1: Four-class classification (%). Sc denotes Score. RespFocus reports five-run means, with SDs in Appendix E. Bold/underlined values mark best/second-best results.

Method	Backbone	Venue	ICBHI 2017			PedResp2026		
			Sc	Se	Sp	Sc	Se	Sp
CE (Moummad & Farrugia, 2023)	CNN6	WASPAA'23	55.19	40.32	70.07	50.00	20.00	80.00
SCL (Moummad & Farrugia, 2023)	CNN6	WASPAA'23	57.57	39.21	75.93	61.92	46.97	76.86
AFT on Mixed (Kim et al., 2023)	AST	NeurIPS W'23	61.80	42.86	80.73	65.31	47.88	82.75
Fine-tuning (Bae et al., 2023)	AST	INTERSPEECH'23	59.52	41.90	77.13	57.08	27.88	<u>86.27</u>
Patch-Mix CL (Bae et al., 2023)	AST	INTERSPEECH'23	62.34	43.02	81.67	<u>68.41</u>	<u>52.12</u>	84.71
Audio-CLAP (Kim et al., 2024b)	CLAP	INTERSPEECH'24	62.73	44.60	80.87	65.24	49.70	80.78
BTS (Kim et al., 2024b)	CLAP	INTERSPEECH'24	63.56	45.71	81.40	64.83	48.48	81.18
DAT (Kim et al., 2024a)	AST	ICASSP'24	59.84	42.54	77.13	64.13	43.94	84.31
SG-SCL (Kim et al., 2024a)	AST	ICASSP'24	61.68	43.49	79.87	64.73	51.82	77.65
M2D-X (Niizumi et al., 2024)	M2D ViT	TASLP'24	63.31	45.08	81.53	62.50	46.97	78.04
MVST (He et al., 2024)	AST	ICASSP'24	<u>66.56</u>	<u>51.11</u>	82.00	54.44	28.48	80.39
Low-rank (Dong et al., 2025a)	LoRA	ICASSP'25	60.68	36.03	**85.33**	56.76	16.67	**96.86**
Single-view (Dong et al., 2025b)	AST	BSPC'25	66.03	49.52	82.53	51.75	38.79	64.71
RespFocus	Dual-stream	This work	**69.55**	**56.03**	<u>83.07</u>	**72.91**	**68.18**	77.65

RespFocus also outperforms the five adapted MIL and temporal baselines in mean Score, macro-AUROC, and macro-F1 on both datasets (Table 7). The strongest baselines by Score are AnchorMoE (Xie et al., 2026) on ICBHI and RRT-MIL (Tang et al., 2024) on PedResp2026. RespFocus exceeds their Scores by 11.91 and 19.85 percentage points, respectively. Compared with RRT-MIL, the strongest macro-F1 baseline on both datasets, RespFocus raises macro-F1 from 0.371 to 0.625 on ICBHI and from 0.372 to 0.744 on PedResp2026. These gains show that the advantage extends to performance averaged equally across the four classes. Figure 3 shows the variation across five runs, with RespFocus Score SDs of 4.92 points on ICBHI and 2.42 on PedResp2026. We next examine the design choices within RespFocus through ablations, complementing these comparisons of complete classification pipelines.

Table 2: Progressive ICBHI 2017 four-class ablation (%, five-run mean ± SD).

Variant	TSCA	Evidence	Binary pair	Asym. pooling	Sp (%)	Se (%)	Score (%)
Flat AST classifier	×	×	×	×	76.40 ± 6.35	43.17 ± 4.07	59.79 ± 4.47
Dual-stream global pooling	✓	×	×	×	79.40 ± 5.67	45.40 ± 3.48	62.40 ± 4.11
Evidence pathway + flat four-way	✓	✓	×	×	81.07 ± 5.91	48.25 ± 3.52	64.66 ± 4.28
Mean pooling	✓	✓	✓	×	82.67 ± 5.48	50.95 ± 3.14	66.81 ± 2.85
Full RespFocus	✓	✓	✓	✓	83.07 ± 8.50	56.03 ± 4.07	69.55 ± 4.92

5.2 Architectural ablations

Adding temporal encoding, evidence guidance, and subtype decisions progressively improves ICBHI classification (Table 2). The mean ICBHI Score rises from 59.79% for a flat AST classifier to 62.40% with dual-stream TSCA, 64.66% with the evidence pathway, and 66.81% with binary-pair decoding and mean pooling. The last comparison holds the remaining architecture and training protocol fixed: replacing mean pooling in both subtype branches with crackle max pooling and wheeze top-k pooling raises Score to 69.55%, a 2.74-point gain. The pooling change raises sensitivity from 50.95% to 56.03%, while specificity changes from 82.67% to 83.07%. Most of the Score increase therefore comes from correct abnormal-subtype predictions. Since this sensitivity metric rewards the correct Crackles, Wheezes, or Both category, the result links the aggregation change to recognition of the sounds present within abnormal cycles. Additional encoder and pooling-ratio comparisons appear in Appendix F.

5.3 Functional role of temporal evidence

The shared evidence distinguishes abnormal cycles and contributes to feature refinement. Its cycle summary r_c achieves an AUROC of $(83.51 \pm 0.48)\%$ for separating abnormal from normal cycles (Table 11(a)). To test its role in refinement, we perturb the shared guidance scores a_t at a fixed checkpoint (Table 10). Replacing the scores with the recording mean removes temporal variation and reduces Score from 70.02% to 65.78%. Removing guidance lowers Score to 62.63%. Within-cycle shuffling preserves each cycle's score distribution but changes its temporal assignment, yielding a Score of 68.28%. These decreases show that the classifier uses the shared guidance, including its temporal arrangement.

Matched feature-space interventions then assess the contribution of the selected representations. Under equal token budgets, suppressing high-evidence tokens reduces mean ICBHI Score by 11.13 percentage points, compared with 4.81 for random tokens and 6.23 for energy-matched random tokens. Retaining only high-evidence tokens yields a Score of 65.12%, versus 55.17% for random retention (Table 12). The larger suppression effect and stronger retention performance jointly support the evidence ranking as a way to identify features that contribute to classification.

Figure 4 illustrates complementary interventions on waveform windows ranked by subtype evidence. On the left, normalized crackle evidence peaks near a brief transient, while wheeze evidence remains elevated over a longer interval. The highlighted windows show where high-evidence and random suppression modify the waveforms. On the right, deleting high-evidence windows causes larger drops in subtype probability than random or energy-matched deletion at partial budgets. Retaining high-evidence windows also yields higher subtype probabilities at small and intermediate budgets. The curves converge at the full-cycle budget, when all strategies delete or retain the same samples. These examples connect the temporal evidence profiles to the prediction changes caused by selecting different portions of the waveform. Appendix H details the waveform protocol.

Figure 5 illustrates two cycles from a wheeze-dominant case. RespFocus correctly predicts Cycle 2 as Both but also assigns Both to the Wheezes-only Cycle 1. The extra crackle prediction propagates into the illustrative text panel. Appendix I provides a detailed cycle-level example.

6 Discussion

RespFocus connects weakly supervised respiratory sound classification with inspection of the temporal evidence used for each decision. The classification results (Table 1) and pooling ablation (Table 2) support using sound morphology to guide how evidence is accumulated. This allows brief

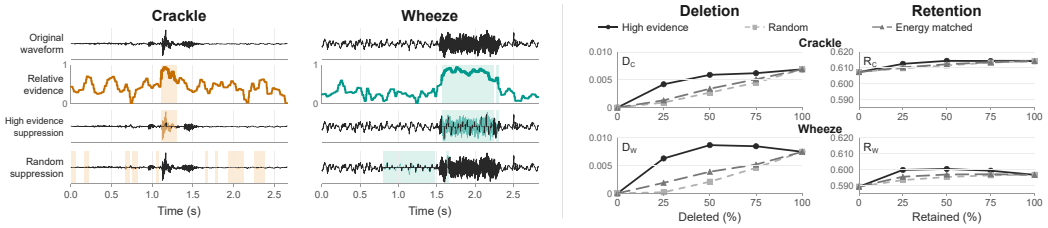


Figure 4: Illustrative waveform interventions with within-cycle min–max-normalized evidence (left). Right panels show deletion $D_q(b) = p_q(x) - p_q(x_{\text{delete},b})$ and retention $R_q(b) = p_q(x_{\text{retain},b})$ for subtype $q \in \{C, W\}$ and cycle-duration fraction b.

crackle responses and longer wheeze activity to contribute through different aggregation rules, while compositional decoding represents their co-occurrence within a cycle. The matched interventions show that the evidence ranking identifies representations on which predictions depend (Table 12). Because these representations retain their temporal indices, the predicted category can be examined alongside the responses that produced it. This gives a listener a concrete starting point for reviewing a prediction and investigating an incorrectly assigned subtype. PedResp2026 provides a new multi-center pediatric setting for studying this connection, with adjudicated recordings from routine care and evaluation on held-out children. The optional text interface conditions generation on acoustic predictions, subtype probabilities, and cycle intervals (Appendix B.3). This conditioning provides an explicit reference for generation and can reduce the risk of language model hallucinations. The extra crackle statement in Figure 5 reflects an error inherited from the acoustic model.

Limitations. PedResp2026 remains limited in size, and future work will expand the cohort to include more patients and a wider range of respiratory sounds. Expert event annotation is also labor-intensive, particularly for brief crackles that require millisecond-scale temporal precision. Learning from cycle-level labels reduces this annotation burden, although evaluating event localization still requires fine-grained reference annotations. The current patient-disjoint split leaves generalization to unseen hospitals and devices open. Future evaluations should test these acquisition-domain shifts (Koh et al., 2021). Evaluation of the text interface should distinguish errors propagated from the acoustic model from unsupported content introduced during text generation (Appendix I). It should also assess how conditioning affects hallucination risk and whether the generated text helps listeners review recordings.

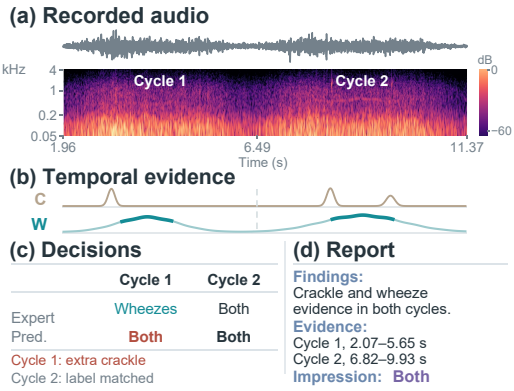


Figure 5: Two-cycle illustration of a wheeze-dominant case.

7 Conclusion

We presented RespFocus, a weakly supervised framework that uses MIL to learn temporal evidence from cycle-level labels and integrate it into respiratory sound classification. Shared abnormal evidence guides separate crackle and wheeze branches, whose pooling rules preserve brief peaks and combine distributed responses before compositional decoding determines the cycle category. Evaluation on ICBHI 2017 and PedResp2026, our new multicenter pediatric cohort, shows improved classification, while matched interventions support the contribution of the learned evidence to predictions. These findings support learning temporal evidence as part of the classifier, connecting each predicted sound category to the temporal responses used to produce it.

AI USE STATEMENT

We used GPT-6 Astra for language polishing, literature search, manuscript-structure review, and checks of numerical consistency and claim–evidence alignment. The authors take full responsibility for the manuscript's content, including its text, references, and scientific claims.

ETHICS STATEMENT

The PedResp2026 study received institutional ethics approval. Written informed consent was obtained from legal guardians, with child assent where appropriate. The data were de-identified before analysis to protect participant privacy. We plan to release PedResp2026 through controlled access upon acceptance to support reproducibility and further research.

REPRODUCIBILITY STATEMENT

Our code and dataset access information are available at `https://anonymous.4open.science/r/RespFous-31C6/`. Dataset splits and annotation procedures are described in Appendix A. Training settings and threshold selection are detailed in Appendix C.

REFERENCES

Sangmin Bae, June-Woo Kim, Won-Yang Cho, Hyerim Baek, Soyoun Son, Byungjo Lee, Changwan Ha, Kyongpil Tae, Sungnyun Kim, and Se-Young Yun. Patch-mix contrastive learning with audio spectrogram transformer on respiratory sound classification. In *Proceedings of Interspeech 2023*, pp. 5436–5440. ISCA, 2023. doi: 10.21437/Interspeech.2023-1426. URL https://doi.org/10.21437/Interspeech.2023-1426.

Dalal Bardou, Kun Zhang, and Sayed Mohammad Ahmad. Lung sounds classification using convolutional neural networks. *Artificial Intelligence in Medicine*, 88:58–69, 2018. doi: 10.1016/j.artmed.2018.04.008. URL https://doi.org/10.1016/j.artmed.2018.04.008.

Yi Chang, Zhao Ren, Thanh Tam Nguyen, Wolfgang Nejdl, and Björn W. Schuller. Example-based explanations with adversarial attacks for respiratory sound analysis. In *Proceedings of Interspeech 2022*, pp. 4003–4007. ISCA, 2022. doi: 10.21437/Interspeech.2022-11355. URL https://www.isca-archive.org/interspeech_2022/chang22h_interspeech.html.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23ag.html.

Xiwen Chen, Peijie Qiu, Wenhui Zhu, Huayu Li, Hao Wang, Aristeidis Sotiras, Yalin Wang, and Abolfazl Razi. TimeMIL: Advancing multivariate time series classification via a time-aware multiple instance learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7190–7206. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24af.html.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4348–4360, 2022. doi: 10.18653/v1/2022.findings-emnlp.319. URL https://aclanthology.org/2022.findings-emnlp.319/.

Thomas G. Dietterich, Richard H. Lathrop, and Tomas Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71, 1997. doi: 10.1016/S0004-3702(96)00034-3. URL https://doi.org/10.1016/S0004-3702(96)00034-3.

Heinrich Dinkel, Mengyue Wu, and Kai Yu. Towards duration robust weakly supervised sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:887–900, 2021. doi: 10.1109/TASLP.2021.3054313. URL https://doi.org/10.1109/TASLP.2021.3054313.

Gaoyang Dong, Yufei Shen, Jianhong Wang, Shunwang Xie, Minghui Zhang, and Ping Sun. Low-rank adaptation method for respiratory sound classification: A necessary road towards large models. In *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025a. doi: 10.1109/ICASSP49660.2025.10889965. URL https://ieeexplore.ieee.org/document/10889965.

Gaoyang Dong, Yufei Shen, Jianhong Wang, Mingli Zhang, Ping Sun, and Minghui Zhang. Respiratory sounds classification by fusing the time-domain and 2D spectral features. *Biomedical Signal Processing and Control*, 107:107790, 2025b. doi: 10.1016/j.bspc.2025.107790. URL https://doi.org/10.1016/j.bspc.2025.107790.

Joseph Early, Gavin Cheung, Kurt Cutajar, Hanting Xie, Jas Kandola, and Niall Twomey. Inherently interpretable time series classification via multiple instance learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fb9d01fb202e360b2f78510c47e0fa3e-Abstract-Conference.html.

Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit Jain. RespireNet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 527–530. IEEE, 2021. doi: 10.1109/EMBC46164.2021.9630091. URL https://doi.org/10.1109/EMBC46164.2021.9630091.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Proceedings of Interspeech 2021*, pp. 571–575. ISCA, 2021. doi: 10.21437/Interspeech.2021-698. URL https://doi.org/10.21437/Interspeech.2021-698.

Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. SSAST: Self-supervised audio spectrogram transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10699–10709, 06 2022. doi: 10.1609/aaai.v36i10.21315. URL https://ojs.aaai.org/index.php/AAAI/article/view/21315.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech 2020*, pp. 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015. URL https://www.isca-archive.org/interspeech_2020/gulati20_interspeech.html.

Wentao He, Yuchen Yan, Jianfeng Ren, Ruibin Bai, and Xudong Jiang. Multi-view spectrogram transformer for respiratory sound classification. In *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8626–8630. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10445825. URL https://doi.org/10.1109/ICASSP48485.2024.10445825.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Jaeseok Jang and Hyuk-Yoon Kwon. TAIL-MIL: Time-aware and instance-learnable multiple instance learning for multivariate time series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17582–17589, 2025. doi: 10.1609/aaai.v39i17.33933. URL https://ojs.aaai.org/index.php/AAAI/article/view/33933.

June-Woo Kim, Chihyeon Yoon, Miika Toikkanen, Sangmin Bae, and Ho-Young Jung. Adversarial fine-tuning using generated respiratory sound to address class imbalance. In *NeurIPS 2023 Workshop on Deep Generative Models for Health*, 2023. URL https://openreview.net/forum?id=z1AVG5LDQ7.

June-Woo Kim, Sangmin Bae, Won-Yang Cho, Byungjo Lee, and Ho-Young Jung. Stethoscope-guided supervised contrastive learning for cross-domain adaptation on respiratory sound classification. In *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1431–1435. IEEE, 2024a. doi: 10.1109/ICASSP48485.2024.10447734. URL https://doi.org/10.1109/ICASSP48485.2024.10447734.

June-Woo Kim, Miika Toikkanen, Yera Choi, Seoung-Eun Moon, and Ho-Young Jung. BTS: Bridging text and sound modalities for metadata-aided respiratory sound classification. In *Proceedings of Interspeech 2024*, pp. 1690–1694. ISCA, 2024b. doi: 10.21437/Interspeech.2024-492. URL https://www.isca-archive.org/interspeech_2024/kim24f_interspeech.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. doi: 10.1109/TASLP.2020.3030497. URL https://doi.org/10.1109/TASLP.2020.3030497.

Ilyass Moummad and Nicolas Farrugia. Pretraining respiratory sound representations using metadata and contrastive learning. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5. IEEE, 2023. doi: 10.1109/WASPAA58266.2023.10248130. URL https://doi.org/10.1109/WASPAA58266.2023.10248130.

Phuc X. Nguyen, Deva Ramanan, and Charless C. Fowlkes. Weakly-supervised action localization with background modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5502–5511, 2019. doi: 10.1109/ICCV.2019.00560. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Nguyen_Weakly-Supervised_Action_Localization_With_Background_Modeling_ICCV_2019_paper.html.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Masked modeling duo: Towards a universal audio pre-training framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2391–2406, 2024. doi: 10.1109/TASLP.2024.3389636. URL https://doi.org/10.1109/TASLP.2024.3389636.

Hans Pasterkamp, Paul L. P. Brand, Mark Everard, Luis Garcia-Marcos, Hasse Melbye, and Kostas N. Priftis. Towards the standardisation of lung sound nomenclature. *European Respiratory Journal*, 47(3):724–732, 2016. doi: 10.1183/13993003.01132-2015. URL https://doi.org/10.1183/13993003.01132-2015.

Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-TALC: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision*, pp. 588–607, 2018. doi: 10.1007/978-3-030-01225-0_35. URL https://doi.org/10.1007/978-3-030-01225-0_35.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17627–17643. PMLR, 2022. URL https://proceedings.mlr.press/v162/peng22a.html.

Huan Ren, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Proposal-based multiple instance learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2394–2404, 2023. doi: 10.1109/CVPR52729.2023.00237. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ren_Proposal-Based_Multiple_Instance_Learning_for_Weakly-Supervised_Temporal_Action_Localization_CVPR_2023_paper.html.

Zhao Ren, Thanh Tam Nguyen, and Wolfgang Nejdl. Prototype learning for interpretable respiratory sound analysis. In *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9087–9091. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9747014. URL https://doi.org/10.1109/ICASSP43922.2022.9747014.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 82–91, 2021. doi: 10.1109/ICCV48922.2021.00015. URL https://openaccess.thecvf.com/content/ICCV2021/html/Ridnik_Asymmetric_Loss_for_Multi-Label_Classification_ICCV_2021_paper.html.

Bruno M. Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P. Kahya, Nikša Jakovljevic, Tatjana L. Turukalo, Ioannis M. Vogiatzis, Eleni Perantoni, Evangelos Kaimakamis, Pantelis Natsiavas, Ana Oliveira, Cristina Jácome, Alda Marques, Nicos Maglaveras, Rui Pedro Paiva, Ioanna Chouvarda, and Paulo de Carvalho. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological Measurement*, 40(3): 035001, 2019. doi: 10.1088/1361-6579/ab03ea. URL https://pubmed.ncbi.nlm.nih.gov/30708353/.

Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11343–11352, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tang_Feature_Re-Embedding_Towards_Foundation_Model-Level_Performance_in_Computational_Pathology_CVPR_2024_paper.html.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7433–7442, 2023. doi: 10.1109/CVPR52729.2023.00718. URL https://openaccess.thecvf.com/content/CVPR2023/html/Tanida_Interactive_and_Explainable_Region-Guided_Radiology_Report_Generation_CVPR_2023_paper.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Tao Xie, Zexi Tan, Haoyi Xiao, Mengke Li, Yiqun Zhang, Yang Lu, Cuie Yang, and Yiu-ming Cheung. Anchormoe: Interpretable time series classification via anchor-routed moe. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 5720–5731, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817786. URL https://doi.org/10.1145/3770855.3817786.

Yifei Xin, Dongchao Yang, and Yuexian Zou. Background-aware modeling for weakly supervised sound event detection. In *Proceedings of Interspeech 2023*, pp. 1199–1203. ISCA,

2023. doi: 10.21437/Interspeech.2023-330. URL https://www.isca-archive.org/interspeech_2023/xin23_interspeech.html.

Yuwei Zhang, Tong Xia, Jing Han, Yu Y. Wu, Georgios Rizos, Yang Liu, Mohammed Mosuily, Jagmohan Chauhan, and Cecilia Mascolo. Towards open respiratory acoustic foundation models: Pretraining and benchmarking. In *Advances in Neural Information Processing Systems*, volume 37, pp. 27024–27055, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f803abdcad9de35b45d5a656dade45c-Abstract-Datasets_and_Benchmarks_Track.html.

Table 3: Patient-disjoint partitions and cycle-level class counts of PedResp2026. Patient and recording counts are followed by percentages of their respective totals in parentheses.

	Partition size			Cycle-level class counts			
Split	Patients	Recordings	Cycles	Normal	Crackles	Wheezes	Both
Train	109 (79.56)	394 (79.92)	899	431	340	116	12
Validation	14 (10.22)	49 (9.94)	119	54	50	14	1
Test	14 (10.22)	50 (10.14)	117	51	45	19	2
Total	137 (100.00)	493 (100.00)	1,135	536	435	149	15

Table 4: ICBHI study partitions and cycle-level class counts.

Split	Recordings	Cycles	Normal	Crackles	Wheezes	Both
Train	777	5,815	3,015	1,595	777	428
Validation	90	657	327	217	58	55
Test	53	426	300	52	51	23
Total	920	6,898	3,642	1,864	886	506

A Datasets

A.1 Splits and unit of analysis

PedResp2026 contains 493 recordings from 137 unique children. All recordings from the same child are assigned to a single partition, ensuring patient-disjoint training, validation, and testing. Patients define the partitions, while respiratory cycles are the units of classification. Table 3 summarizes partition sizes and cycle-level class counts.

The ICBHI partition is fixed across the reported experiments, including the additional metrics. The test recordings follow the OPERA held-out test list. Table 4 reports the recording and cycle counts and the cycle-level class distributions for each split.

A.2 Cohort, Acquisition, and Annotation

PedResp2026 comprises 137 unique children, including 80 male (58.39%) and 57 female (41.61%) participants, who contribute 493 recordings from seven source hospitals. Recorded ages range from approximately 0.1 to 13 years. Clinical diagnoses describe the cohort composition and are distinct from the four acoustic classification targets. Figure 6 summarizes recorded ages, patient sex, clinical diagnoses, and source hospitals.

Recordings were acquired in routine, non-acoustically controlled pediatric clinical rooms. The collection protocol covered eight anterior and posterior thoracic sites, including left and right upper and lower chest regions. Each auscultation session lasted at least 30 seconds.

Two physicians first annotated each respiratory cycle. An additional physician from the eight-physician pediatric panel reviewed and adjudicated the labels, and the final adjudicated labels were used for training and evaluation.

The study received institutional ethics approval. Written informed consent was obtained from legal guardians before collection, with child assent where appropriate for age and comprehension. Data were de-identified before analysis, and institution-identifying details are withheld during double-blind review. We plan to release PedResp2026 through controlled access upon acceptance.

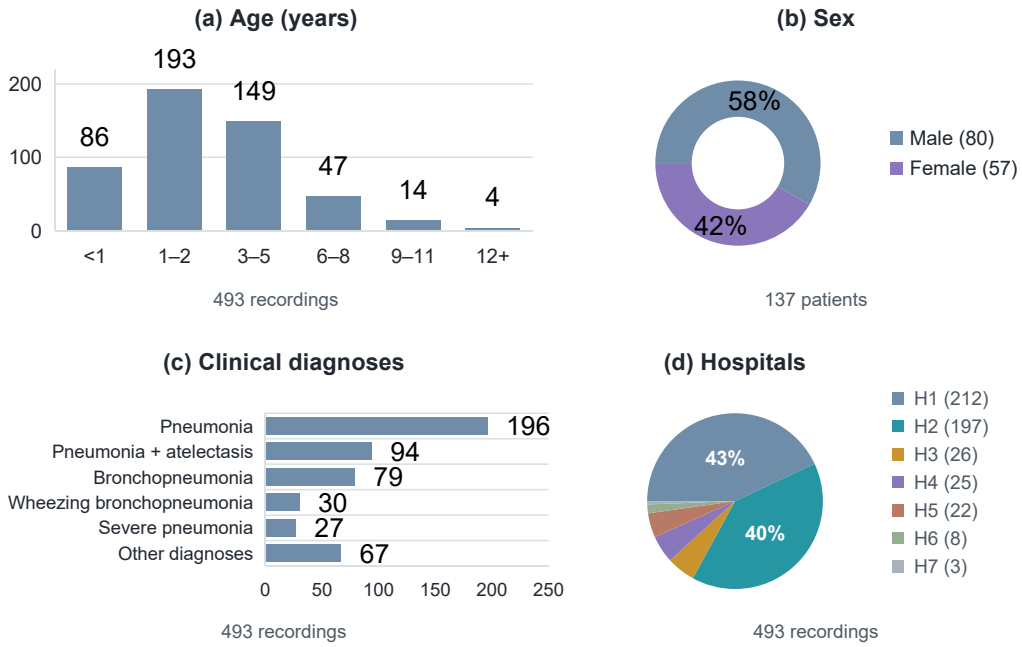


Figure 6: PedResp2026 cohort composition. (a) Recorded ages. (b) Patient sex. (c) The five most frequent recorded diagnoses, with remaining labels grouped as Other. (d) Anonymized source hospitals. Panels (a,c,d) count recordings ($n = 493$), while (b) counts patients ($n = 137$). In (a), integer age ranges denote completed years, and 12+ includes ages 12–13. Legend entries in (b,d) give counts.

B Additional Method Details

B.1 Shared Evidence Computation

The shared logit combines a local evidence score with an auxiliary abnormal cue. For each fused token $h_t \in \mathbb{R}^{512}$, the auxiliary scorer $g_{\text{aux}} = \text{Linear}(512, 4)$ produces four logits. Its implementation channel labels, in order, are Normal, Crackles, Wheezes, and Both. The abnormal-channel index set is $\mathcal{A} = \{1, 2, 3\}$ under zero-based indexing. We compute

$$\begin{aligned} z_t^{\text{loc}} &= f_{\text{loc}}(h_t), \\ m_t &= \log\left(\frac{1}{3}\sum_{j\in\mathcal{A}} \exp([g_{\text{aux}}(h_t)]_j)\right), \\ s_t &= z_t^{\text{loc}} + \lambda_{\text{cue}} m_t, \qquad a_t = \sigma(s_t/\tau_a). \end{aligned} \tag{16}$$

We set $\tau_a = 1.0$ and increase λ_{cue} linearly from 0 to 0.2 over the first eight acoustic-training epochs, then keep it at 0.2. Residual guidance in Section 3.2 uses $\alpha = 0.5$ with stop-gradient enabled. The shared-evidence objective in Section 3.4 is enabled by `binary_pair_aux_presence_head=1`, with `binary_loss_lambda=1.0`. The four auxiliary channel names specify the cue construction. Final categories follow the decoder in Section 3.3.

B.2 Local Refinement

Inspired by the convolutional gating design of Branchformer (Peng et al., 2022), we implement a residual local refinement block for each subtype, with the operations ordered as follows:

$$F_k(X) = X + \gamma W_o(\text{GELU}(\text{DWConv}_k(\text{GLU}(W_e(\text{LN}(X)))))) \,. \tag{17}$$

Here, W_e and W_o are linear projections and γ is a learnable residual scale. The kernel sizes and subtype heads follow Section 3.2. Figure 2(b) shows the corresponding refinement blocks.

B.3 Structured summary and language interface

Medical report generation provides related interfaces between perceptual evidence and text. Region-guided radiology reporting links findings to image regions (Tanida et al., 2023), while semantic rewards encourage factual correctness (Delbrouck et al., 2022). RespFocus uses an optional presentation branch to verbalize acoustic predictions and their temporal references.

The presentation branch in Figure 2(d) uses the structured summary

$$E_c = \{\bar{h}_c, r_c, p_c^C, p_c^W, \hat{y}_c\}, \qquad \bar{h}_c = \frac{1}{|\mathcal{I}_c|} \sum_{t \in \mathcal{I}_c} \tilde{h}_t. \tag{18}$$

The mean guided feature $\bar{h}_c$ complements the support r_c defined in Section 3.4 and the subtype probabilities and label from Section 3.3. The cycle interval I_c supplies start and end times in seconds.

The audio interface takes the fused temporal features after the output projection in Section 3.1, restricted to each cycle: $A_c = [h_t]_{t \in \mathcal{I}_c} \in \mathbb{R}^{|\mathcal{I}_c| \times 512}$ in temporal order. A Q-Former with eight learnable queries, two layers, eight attention heads, and hidden dimension 512 produces eight tokens. An MLP projector maps them to the language-model embedding dimension d_{LM},

$$\begin{aligned} Q_c &= \mathrm{QFormer}(A_c) \in \mathbb{R}^{8 \times 512}, \\ V_c &= \mathrm{MLP}_{\mathrm{proj}}(Q_c) \in \mathbb{R}^{8 \times d_{\mathrm{LM}}}. \end{aligned} \tag{19}$$

MedGemma 1.5 4B-it (`medgemma-1.5-4b-it`) generates text from the projected audio tokens, E_c, and interval metadata. Appendix C gives the adaptation configuration, and Appendix I presents illustrative outputs.

C Implementation and thresholds

Encoder inputs. OPERA-CT uses `encoder-operaCT.ckpt` and a 64-bin log-mel spectrogram from 16-kHz audio, with a 2,048-sample FFT/window, a 512-sample hop, and an 80-dB ceiling. Long sequences use chunks of at most 1,024 audio tokens, with adaptive pooling when the limit is exceeded.

AST uses `MIT/ast-finetuned-audioset-10-10-0.4593` and a 128-bin Kaldi filterbank with a 10-ms frame shift. Normalization uses `mean=-4.2677393`, `std=4.5689974`, and `std_scale=2.0`. Inputs are zero-padded to 512 frames when shorter and truncated to the first 512 frames when longer, without sliding windows. Fusion dimensions are specified in Section 3.1.

Acoustic optimization. AdamW uses an initial learning rate of 3.5×10^{-5} and weight decay 0.08. Training runs for 25 epochs, with five warm-up epochs, cosine decay, and a minimum learning-rate ratio of 0.15. The configured batch size is 10 with two gradient-accumulation steps (reported effective batch size: 20). Experiments use four NVIDIA H100 GPUs and seeds 0–4. For the objective in Section 3.4, ASL uses $\gamma_- = 4$, $\gamma_+ = 1$, and probability clipping $m = 0.05$.

Language-branch adaptation. The acoustic model is initialized from the checkpoint `best.pt`. Only the Q-Former, projector, and MedGemma LoRA parameters are trained. OPERA, AST, TSCA, acoustic heads, and base MedGemma weights are frozen. LoRA (Hu et al., 2022) uses rank 16, scaling parameter 32, dropout 0.05, and `target_modules=all-linear`. Adaptation runs for eight epochs with batch size 2, four gradient-accumulation steps, and learning rate 10^{-4}. SpecAugment, synthetic-negative generation, and synthetic-Both generation are disabled. All parameters are fixed at inference.

Table 5: Validation-selected subtype thresholds. Each entry is (θ_C, θ_W), and decisions use $p_c^q > \theta_q$.

Seed	ICBHI 2017	PedResp2026
0	$(0.55, 0.60)$	$(0.65, 0.65)$
1	$(0.60, 0.60)$	$(0.60, 0.70)$
2	$(0.65, 0.65)$	$(0.65, 0.70)$
3	$(0.55, 0.55)$	$(0.60, 0.70)$
4	$(0.55, 0.60)$	$(0.60, 0.65)$

Threshold selection. We jointly search (θ_C, θ_W) to maximize validation four-class Score. Each threshold ranges from 0.05 to 0.80 in steps of 0.025 on ICBHI, and from 0.20 to 0.80 in steps of 0.05 on PedResp2026. Table 5 lists the selected pairs used for test evaluation.

Table 6: Class-wise F1 over five independent runs (mean ± sample standard deviation), on a 0–1 scale.

Dataset	Normal F1	Crackles F1	Wheezes F1	Both F1
ICBHI 2017	0.8355 ± 0.0728	0.7211 ± 0.0683	0.4891 ± 0.0685	0.4527 ± 0.0543
PedResp2026	0.7226 ± 0.0433	0.6856 ± 0.0327	0.7963 ± 0.0552	0.7733 ± 0.0596

Table 7: Mean test-set performance of adapted MIL and interpretable temporal models. Score and AUROC are percentages, and macro-F1 is on a 0–1 scale. Run-level variation is shown in Figure 3.

Method	Venue	Score (%)	Macro-AUROC (%)	Macro-F1
ICBHI 2017				
TimeMIL	ICML'24	45.92	49.57	0.235
TAIL-MIL	AAAI'25	52.48	50.79	0.287
RRT-MIL	CVPR'24	56.58	60.31	0.371
MILLET	ICLR'24	54.00	52.41	0.294
AnchorMoE	KDD'26	57.64	52.55	0.323
RespFocus	This work	69.55	71.86	0.625
PedResp2026				
TimeMIL	ICML'24	46.97	44.20	0.234
TAIL-MIL	AAAI'25	49.00	51.73	0.226
RRT-MIL	CVPR'24	53.06	62.98	0.372
MILLET	ICLR'24	47.24	48.91	0.211
AnchorMoE	KDD'26	52.39	61.97	0.343
RespFocus	This work	72.91	78.71	0.744

D Additional metrics

For AUROC evaluation, the two cycle-level subtype probabilities define four continuous class scores,

$$\begin{aligned} \pi_c^N &= (1-p_c^C)(1-p_c^W), \quad \pi_c^C = p_c^C(1-p_c^W), \\ \pi_c^W &= (1-p_c^C)p_c^W, \qquad\quad \pi_c^B = p_c^C p_c^W . \end{aligned} \tag{20}$$

Here, π_c^C denotes the Crackles-only score and y_c the reference four-class label. Macro-AUROC averages the one-vs-rest AUROCs,

$$\mathrm{MacroAUROC} = \frac{1}{4} \sum_{q \in \{N,C,W,B\}} \mathrm{AUROC}(\{\mathbf{1}[y_c = q]\}_c, \{\pi_c^q\}_c) . \tag{21}$$

AUROC uses continuous scores, while F1 uses the hard predictions from Section 3.3. Table 7 reports the macro metrics, and Table 6 provides the class-wise F1 breakdown over five seeds. In Table 6, Both has the lowest class-wise F1 on ICBHI. Its PedResp2026 estimate is based on two positive test cycles (Table 3).

D.1 Adapted MIL and temporal baselines

Table 7 reports the mean test-set performance of RespFocus and the adapted baselines on ICBHI 2017 and PedResp2026.

Table 8: Full performance comparison on ICBHI 2017 and PedResp2026. We report Score, sensitivity (Se), and specificity (Sp), in percent. Ours reports mean ± standard deviation over five independent runs. Best and second-best means in each dataset–task block are bold and underlined. Audio-CLAP four-class Score is recomputed from its reported Sp and Se. PedResp2026 baselines use the study protocol.

Method	Backbone	Venue	ICBHI 2017			PedResp2026		
			Score (%)	Se (%)	Sp (%)	Score (%)	Se (%)	Sp (%)
Four-class								
CE	CNN6	WASPAA'23	55.19 ± 0.55	40.32 ± 2.99	70.07 ± 3.97	50.00 ± 5.04	20.00 ± 11.46	80.00 ± 1.64
SCL	CNN6	WASPAA'23	57.57 ± 0.80	39.21 ± 1.91	75.93 ± 2.31	61.92 ± 2.96	46.97 ± 3.39	76.86 ± 6.99
AFT on Mixed	AST	NeurIPS W'23	61.80 ± 0.47	42.86 ± 1.25	80.73 ± 1.01	65.31 ± 0.93	47.88 ± 3.65	82.75 ± 3.22
Fine-tuning	AST	INTERSPEECH'23	59.52 ± 0.89	41.90 ± 2.20	77.13 ± 3.35	57.08 ± 5.76	27.88 ± 13.15	$\underline{86.27} \pm 6.20$
Patch-Mix CL	AST	INTERSPEECH'23	62.34 ± 0.61	43.02 ± 2.83	81.67 ± 3.83	$\underline{68.41} \pm 1.71$	$\underline{52.12} \pm 6.99$	84.71 ± 6.26
Audio-CLAP	CLAP	INTERSPEECH'24	62.73 ± 0.38	44.60 ± 3.78	80.87 ± 3.33	65.24 ± 4.75	49.70 ± 9.43	80.78 ± 4.68
BTS	CLAP	INTERSPEECH'24	63.56 ± 0.80	45.71 ± 2.67	81.40 ± 2.55	64.83 ± 3.39	48.48 ± 5.77	81.18 ± 3.56
DAT	AST	ICASSP'24	59.84 ± 1.25	42.54 ± 5.40	77.13 ± 7.20	64.13 ± 1.79	43.94 ± 4.15	84.31 ± 3.67
SG-SCL	AST	ICASSP'24	61.68 ± 1.61	43.49 ± 5.92	79.87 ± 8.89	64.73 ± 1.54	51.82 ± 8.73	77.65 ± 9.26
M2D-X	M2D ViT	TASLP'24	63.31 ± 0.22	45.08 ± 1.03	81.53 ± 1.02	62.50 ± 2.52	46.97 ± 3.86	78.04 ± 4.88
MVST	AST	ICASSP'24	$\underline{66.56} \pm 2.71$	$\underline{51.11} \pm 5.19$	82.00 ± 4.60	54.44 ± 4.27	28.48 ± 6.10	80.39 ± 8.08
Low-rank	LoRA	ICASSP'25	60.68 ± 1.20	36.03 ± 2.22	$\mathbf{85.33} \pm 2.07$	56.76 ± 4.07	16.67 ± 8.83	$\mathbf{96.86} \pm 1.75$
Single-view	AST	BSPC'25	66.03 ± 2.87	49.52 ± 3.91	82.53 ± 4.94	51.75 ± 2.85	38.79 ± 10.69	64.71 ± 8.43
Ours	Dual-stream	This work	$\mathbf{69.55} \pm 4.92$	$\mathbf{56.03} \pm 4.07$	$\underline{83.07} \pm 8.50$	$\mathbf{72.91} \pm 2.42$	$\mathbf{68.18} \pm 5.46$	77.65 ± 8.50
Two-class								
DAT	AST	ICASSP'24	67.06 ± 1.28	56.98 ± 7.41	77.13 ± 7.20	70.19 ± 4.35	56.06 ± 12.31	84.31 ± 3.67
SG-SCL	AST	ICASSP'24	$\underline{68.90} \pm 1.45$	$\underline{57.94} \pm 8.96$	79.87 ± 8.89	70.49 ± 3.61	$\underline{63.33} \pm 10.46$	77.65 ± 9.26
Fine-tuning	AST	INTERSPEECH'23	66.74 ± 1.15	56.35 ± 1.68	77.13 ± 3.35	72.68 ± 0.67	59.09 ± 5.36	$\mathbf{86.27} \pm 6.20$
Patch-Mix CL	AST	INTERSPEECH'23	68.69 ± 0.61	55.71 ± 2.83	$\underline{81.67} \pm 3.83$	$\underline{72.96} \pm 1.53$	61.21 ± 6.48	$\underline{84.71} \pm 6.26$
Ours	Dual-stream	This work	$\mathbf{72.96} \pm 11.16$	$\mathbf{62.86} \pm 15.10$	$\mathbf{83.07} \pm 8.50$	$\mathbf{74.58} \pm 3.64$	$\mathbf{71.52} \pm 4.96$	77.65 ± 8.50

Table 9: Additional one-factor ablations on ICBHI 2017 four-class classification. Metrics are percentages.

Group	Variant	Sp (%)	Se (%)	Score (%)
Acoustic view	AST-only single-stream	76.40	43.17	59.79
Acoustic view	OPERA-only single-stream	77.40	43.81	60.60
Acoustic view	Dual-stream with concatenation	78.13	44.60	61.37
Acoustic view	Dual-stream with TSCA (baseline)	79.40	45.40	62.40
Pooling	Wheeze top-k, $\rho_W = 0.50$	81.33	50.63	65.98
Pooling	Wheeze top-k, $\rho_W = 0.25$	82.67	52.86	67.76
Pooling	Wheeze top-k, $\rho_W = 0.35$ (full)	83.07	56.03	69.55

E Full Classification Benchmark

Table 8 provides the complete four-class and binary benchmark results. Binary evaluation groups Crackles, Wheezes, and Both as Abnormal.

F Additional Ablations

Table 9 compares single-stream encoders, concatenation versus TSCA, and wheeze pooling ratios. Among the tested ratios, $\rho_W = 0.35$ yields the highest reported Score.

G Operational Evidence Perturbations

Table 10 evaluates changes to a_t with one acoustic checkpoint and fixed thresholds, subtype heads, pooling, and decoder. No guidance sets $a_t = 0$. Uniform evidence uses the recording mean. Within-cycle shuffling permutes the scores inside each cycle and averages ten permutations. Relative to the 70.02% baseline in Table 10, the Score decreases by 7.39, 4.24, and 1.73 points, respectively. Removing guidance has a larger effect than disrupting its within-cycle alignment.

Table 10: Single-checkpoint operational evidence-shuffle ablation on ICBHI 2017. Only the abnormal evidence used for residual guidance is modified at inference. Metrics are percentages, and Δ is the Score change in percentage points.

Evidence condition	Sp (%)	Se (%)	Score (%)	Δ
Original evidence with temporal alignment	77.33	62.70	70.02	–
No evidence guidance	73.67	51.59	62.63	-7.39
Uniform evidence	76.00	55.56	65.78	-4.24
Within-cycle shuffled evidence	76.97	59.60	68.28	-1.73

Table 11: Shared-evidence diagnostics and token suppression on ICBHI 2017. (a) CycleHitRate measures within-cycle high-evidence support, while AUROC evaluates Normal-versus-Abnormal ranking. (b) Suppression uses a matched cycle-level token budget. Values are mean $\pm$ SD in percent, and Δ is the Score change in percentage points.

(a) Shared-evidence diagnostics

Variant	CycleHitRate (%)	Abnormal-presence AUROC (%)
Evidence head w/o guidance	56.35 ± 4.16	75.82 ± 1.54
w/o abnormal-presence guidance	98.10 ± 2.95	67.49 ± 0.50
Full RespFocus	82.86 ± 1.20	83.51 ± 0.48

(b) Token suppression

Perturbation	Sp (%)	Se (%)	Score (%)	Δ
Unperturbed input	83.07 ± 8.50	56.03 ± 4.07	69.55 ± 4.92	–
Remove high-evidence tokens	78.27 ± 6.87	38.57 ± 4.03	58.42 ± 5.18	-11.13
Remove random tokens	81.07 ± 6.44	48.41 ± 3.93	64.74 ± 4.94	-4.81
Remove low-evidence tokens	82.73 ± 6.02	52.54 ± 3.39	67.64 ± 4.51	-1.91

H EVIDENCE DIAGNOSTICS AND FEATURE-SPACE INTERVENTIONS

Table 11(a) reports an abnormal-presence AUROC of $(83.51 \pm 0.48)\%$ for the shared evidence. This metric evaluates r_c against Normal-versus-Abnormal cycle labels and is distinct from four-class macro-AUROC. The variant without abnormal-presence guidance has a higher CycleHitRate but lower AUROC, illustrating that widespread activation within abnormal cycles need not be discriminative. Table 11(b) tests whether classification depends on the selected evidence. Removing high-evidence tokens reduces Score by 11.13 points, compared with 4.81 for random tokens and 1.91 for low-evidence tokens under the matched suppression budget. The larger reduction supports the functional relevance of the evidence ranking to classification. The extended comparison in Table 12 includes an energy-matched random control, whose 6.23-point reduction provides a comparison beyond acoustic energy alone.

Table 12 compares token suppression and retention under a matched cycle-level budget. Suppression uses the same feature-space replacement for high-evidence, random, low-evidence, and energy-matched random selections. The energy-matched control samples from the same cycle and log-mel energy bin as each high-evidence token. Section 5.3 discusses the resulting Score changes.

Waveform protocol for the illustrative interventions. For the illustrative profiles in Figure 4, we rank windows using the subtype evidence $e_q(t) = \sigma(u_t^q)$, $q \in \{C, W\}$, from the unperturbed input. The logits u_t^q are the subtype-head outputs before cycle pooling (`cycle_binary_pair_token_logits`). A cycle of duration T_c is partitioned into N_c equal-width windows, with a budget of K_c^{int} windows:

$$N_c = \max\left(4, 4\left\lceil \frac{T_c}{4\Delta t} \right\rceil\right), \qquad K_c^{\text{int}} = \text{round}(bN_c), \qquad \Delta t \approx 32\,\text{ms}. \tag{22}$$

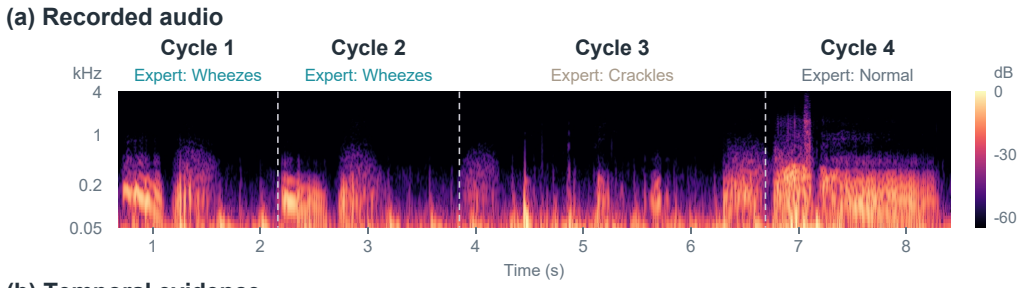
Here, b denotes the cycle-duration budget fraction. Window scores are overlap-duration-weighted averages of the token scores. The highest-scoring windows are selected, with the ranking fixed throughout the interventions. Suppression zeros the selected windows in the raw waveform before OPERA and AST preprocessing, with a 2-ms cosine transition on each side. The modified full recording is passed through both frontends and the downstream model, preserving recording length, cycle boundaries, and audio outside the target cycle.

Table 12: Extended temporal evidence interventions on ICBHI 2017. Suppression modes use a matched cycle-level token budget and feature-space replacement. The energy-matched control matches log-mel energy. Metrics are mean ± standard deviation over five runs, in percent. Δ is the Score change in percentage points.

Temporal evidence perturbation	Sp (%)	Se (%)	Score (%)	Δ
Unperturbed input	83.07 ± 8.50	56.03 ± 4.07	69.55 ± 4.92	–
Suppress high-evidence tokens	78.27 ± 6.87	38.57 ± 4.03	58.42 ± 5.18	-11.13
Suppress random tokens	81.07 ± 6.44	48.41 ± 3.93	64.74 ± 4.94	-4.81
Suppress energy-matched random tokens	80.13 ± 6.52	46.51 ± 3.95	63.32 ± 4.88	-6.23
Suppress low-evidence tokens	82.73 ± 6.02	52.54 ± 3.39	67.64 ± 4.51	-1.91
Retain only high-evidence tokens	80.40 ± 6.35	49.84 ± 3.74	65.12 ± 4.73	-4.43
Retain random tokens	74.93 ± 7.21	35.40 ± 4.30	55.17 ± 5.42	-14.38

I Qualitative cases and optional text presentation

Figure 7 illustrates the relationship between the recorded audio, subtype evidence pooling, and cycle-level decisions for the first four annotated respiratory cycles of a recording. The cycles span 0.679–8.407 s, with reference labels of Wheezes, Wheezes, Crackles, and Normal, respectively. The spectrogram is computed from the recording. The crackle branch applies max pooling to its temporal logits, whereas the wheeze branch averages the highest 35% of temporal logits before applying the sigmoid function. The darker wheeze segments indicate the selected portions of the illustrative profile. A Crackles prediction requires the pooled crackle probability to exceed its decision threshold. For the four displayed cycles, the supplied crackle probabilities are below 0.55 and the supplied wheeze probabilities are above 0.60. Each displayed prediction includes the corresponding annotated cycle interval. As shown, the decisions agree with the reference labels for Cycles 1 and 2. Cycle 3 illustrates a subtype error, and Cycle 4 illustrates a false-positive abnormal prediction. The example illustrates the decision rule and its errors.



(c) Cycle decisions

Thresholds: $pC > 0.55$, $pW > 0.60$

pC	0.517	0.507	0.518	0.525
pW	0.654	0.649	0.628	0.669
Pred.	**Wheezes**	**Wheezes**	**Wheezes**	**Wheezes**
Time (s)	0.679–2.164	2.164–3.850	3.850–6.693	6.693–8.407

Figure 7: Cycle-level illustration for the first four cycles of `164_1b1_Ll_sc_Meditron`. **(a)** Recorded spectrogram with cycle boundaries and labels. **(b)** Illustrative subtype evidence. Darker wheeze segments indicate tokens selected for pooling. **(c)** Cycle probabilities, time intervals, and predictions. Incorrect predictions are shown in red.